# Approach for Constructing Chatbots About Government Human Resources Policies

**Steve Desilets**†

† **Address to which correspondence should be addressed:**

**Steve.desilets27@gmail.com**

# Agenda

1. Problem Statement
2. Analysis and Interpretation
3. Summarization & Recommendations
4. Executive Summary
5. Learn More & Connect

# Problem Statement

PROBLEM STATEMENT

# Introduction

**Context That Inspired Chatbots About HR Policy at HHS**

Numerous government human resources (HR) offices have recognized that their outdated systems fail to adequately equip HR specialists with the tools necessary for the delivery of high-quality HR customer service to agency staff (ICF Incorporated 2019; USAID 2016). In response, many of these government agencies have recognized the criticality of transforming the way in which they deliver HR services to be more technologically advanced and streamlined (ICF Incorporated 2019; USAID 2016).

Inspired by the operational efficiency and cost savings that these agencies have reaped from their recent investments in revitalized HR support technology, the U.S. Department of Health and Human Services (HHS) Office of Human Resources (OHR) is embarking upon a similar HR service delivery technological transformation (HHS OHR 2023).

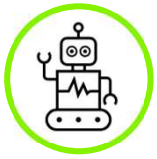
Building on this enthusiasm, we researched the role that natural language processing (NLP) – based chatbots could play in empowering HHS OHR staff to rapidly obtain answers to questions regarding HHS HR policy, so that HR specialists can more rapidly and accurately address HHS staff's HR needs.

# Problem Statement

Below, we provide the objectives that we set out to achieve when constructing chatbots about HR policy at HHS.

## Leverage NLP Methods to Advance the Accuracy and Efficiency of HR Service Delivery at HHS

We hope to empower HHS to increase the operational efficiency and accuracy of OHR's HR service delivery. We hope to achieve this objective by identifying the best way to construct HR policy question-answering chatbots that can support HHS HR Specialists, so they can better serve HHS staff and support the Department's mission.



## Identify Chatbot Development Best Practices that Could Help Other Government HR Offices

We hope that our research, which identifies the chatbot model types and model design choices best suited for OHR, may allow HR Offices at other government agencies to successfully implement similar HR Policy Question-Answering Chatbots using our lessons learned, recommendations, and chatbot development code.

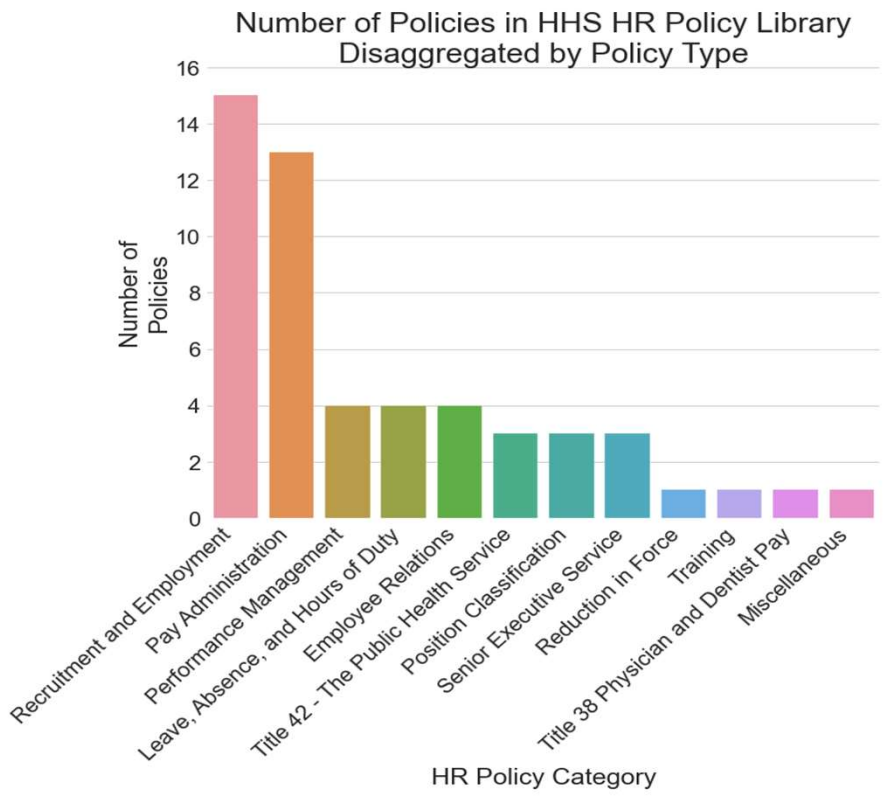
# Analysis and Interpretation

ANALYSIS AND INTERPRETATION

# Data

The dataset leveraged to train the chatbots in this study consists of the HR policy documents included within the HHS Human Resources Policy Library (HHS 2024). The Python-generated visualization and summary statistics below convey important information about the corpus. Given the large size of this policy library, we were confident that the chatbots would have a rich corpus from which to extract information.

## HHS HR Policy Library Documents By Topic




## Corpus Summary Statistics

**223,622**
Words In the Corpus

**53**
Policy Documents in the Corpus

**12**
Categories of HHS HR Policies

ANALYSIS AND INTERPRETATION

# Methodology

This slide summarizes the methodology leveraged to construct each of the chatbots focused on answering questions about HHS HR policies.

## 1

### Sentence-Based Transformer Model

This pre-trained Sentence Transformer Model, which was pre-trained via 215 million question-answer pairs, encodes each sentence of the corpus, as well as the user's question, as 384-dimensional vectors. Then, the model calculates the cosine similarity between the question vector and each of the corpus sentence vectors and returns the corpus sentence associated with the highest cosine similarity with the user's question (Espejel 2022).

## 2

### Fine-Tuned GPT2 Model

This chatbot was created by fine tuning the ChatGPT-2 model in Python using sample training questions about coral reef fish that were entirely separate from our testing questions (OpenAI 2019). After finishing this fine-tuning process, this large language model encoded and responded to each user question.

## 3

### TF-IDF Cosine Similarity Model

This chatbot first tokenizes the corpus by sentence. It then conducts data wrangling on each sentence in the corpus and in the user's question - including word tokenization, lemmatization, transformation to lowercase, and punctuation removal. The chatbot applies TF-IDF vectorization to each sentence, calculates the cosine similarity between each of the sentences, and returns the sentence from the corpus with the highest cosine similarity to the user's question (Kulkarni 2020).

## 4

### DistilBERT Model

Like its predecessor, the BERT Model, the DistilBERT model is a transformer-based model that was pretrained to complete masked language modeling and next sentence prediction (Sanh et al. 2019). The DistilBERT model has 40% fewer parameters than the BERT model, which allows the DistilBERT model to be much less expensive computationally while preserving 95% of BERT's level of performance. After providing this DistilBERT model with our corpus, the chatbot was ready to answer questions about HHS HR policy.

## 5

### Roberta Model

The fifth chatbot that we leveraged was a version of the Roberta model that was developed for the purpose of answering questions (Chan et al. 2023). The Roberta model is pretrained to complete masked language modeling. However, the training times for Roberta Models are generally longer than the training times for comparable large language models, like DistilBERT. We provided the corpus of text to the Roberta Model, so that it would then be able to answer questions about HHS HR policy.

## 6

### Ensemble Model

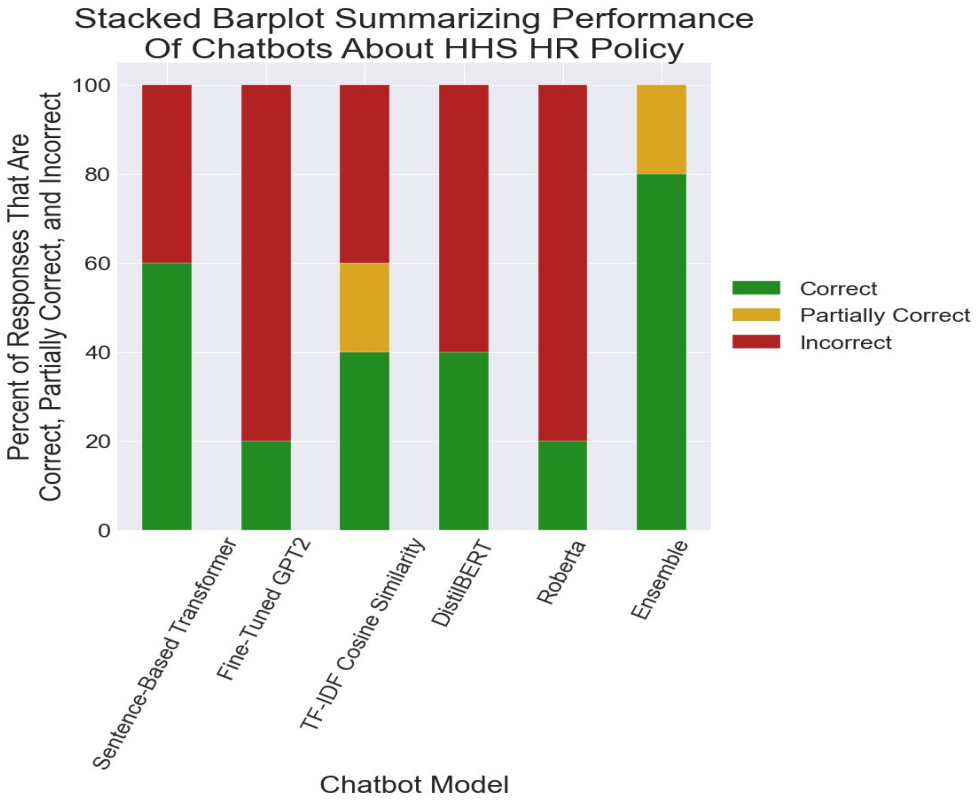
The Ensemble Model leverages a two-part mixed methods design to generate answers to user questions. After completing data wrangling and creating embedding representations of each HR policy document and of the user question, the model returns the three documents with the greatest cosine similarity to the user question. The second component of this mixed model then prompts OpenAI's ChatGPT 3.5 model to generate a response to the user question based on the context provided from these three most relevant HR policy documents. This model aims to harness the all the knowledge within the whole corpus while limiting the computational burden imposed upon the transformer-based model by only loading the most relevant documents into the transformer model (Masri 2023).

# Summarization and Recommendations

# Results – Accuracy and Speed of Chatbot Responses

The visualizations below summarize the results from testing each of our chatbots with 5 questions, so we could measure their response speed and accuracy. The results convey that while three models – the Ensemble Methods Model, the TF-IDF Cosine Similarity Model, and the Fine-Tuned GPT2 Model – all performed well when considering chatbot response times, the Ensemble Methods Model performed most strongly when measuring chatbot response accuracy rates.

## Chatbot Response Accuracy




## Chatbot Response Speed

| Mean Chatbot Response Time to Test Questions | |
|---|---|
| **Chatbot Model** | **Mean Response Time (in seconds)** |
| Sentence-Based Transformer | 180 |
| Fine-Tuned GPT2 | 8 |
| TF-IDF Cosine Similarity | 1 |
| DistilBERT | 1,001 |
| Roberta | 2,280 |
| Ensemble Methods | 1 |

# Results – Chatbot Responses| Part 1

The table below displays the chatbot responses to the test questions. The table's cells are color coded as green, yellow, and red to reflect whether each answer was correct, partially correct, or incorrect, respectively. The target answers for each test question are provided in the appendix.

| Question / Chatbot Response | Question 1 and Responses | Question 2 and Responses | Question 3 and Responses | Question 4 and Responses | Question 5 and Responses |
|---|---|---|---|---|---|
| Question | Name one or more means of handling misconduct | Up to how much money can an individual receive per year from the Student Loan Repayment Program? | Where are employee records, such as annual ratings under a performance appraisal program, held? | Up to how many years can an employee serve overseas in a single country on a tour of duty? | Which types of doctors are eligible for PDP Pay? |
| Sentence-Based Transformer Model Response | examples of misconduct include tardiness, absenteeism, unprofessional or discourteous conduct, damaging or destroying government property, or falsification. | eligible employees may be considered for loan repayment assistance up to $10,000 per calendar year, with a $60,000 lifetime maximum for any individual. | the regulations specify that appraisal periods shall generally be designated so that employees are provided a rating of record annually. | employees may serve up to a maximum of six (6) consecutive years in a single country. | physician, dentist and podiatrist pay (pdp) |
| Fine-Tuned GPT2 Model Response | Labor Department | Up to how much money an individual can receive per year from the Student Loan Repayment Program. | Employee records are held by the Department of Labor under the Employee Retirement Income Security Act (ERISA). Employee records are issued to employees only for performance reasons, and are not issued to employees for performance reasons. | Up to how long an employee can serve overseas in a single country on a tour of duty. | PDP Pay is a payment made to a physician for the performance of a service rendered by a qualified health professional in accordance with the provisions of the Medicare Part D Act. PDP Pay is made available to all publicly funded health care facilities except for emergency room physicians, which are subject to the same payment schedules as the private sector. |

## Findings

+ Every model in this study failed to answer the third question correctly – mostly because they returned information related to performance appraisals rather than return information related to the document storage system – the Employee Performance File (EPF) System

+ The Fine-Tuned GPT2 Model repeatedly returned the test question as its response

# Results – Chatbot Responses| Part 2

The table below displays the chatbot responses to the test questions. The table's cells are color coded as green, yellow, and red to reflect whether each answer was correct, partially correct, or incorrect, respectively. The target answers for each test question are provided in the appendix.

| Question / Chatbot Response | Question 1 and Responses | Question 2 and Responses | Question 3 and Responses | Question 4 and Responses | Question 5 and Responses |
|---|---|---|---|---|---|
| Question | Name one or more means of handling misconduct | Up to how much money can an individual receive per year from the Student Loan Repayment Program? | Where are employee records, such as annual ratings under a performance appraisal program, held? | Up to how many years can an employee serve overseas in a single country on a tour of duty? | Which types of doctors are eligible for PDP Pay? |
| TF-IDF Cosine Similarity Model Response | b. means of handling misconduct<br><br>counseling or verbal warnings should be used by the supervisor when: infractions are minor and infrequent; and the supervisor determines that counseling and/or verbal warnings will likely preclude a recurrence of the misconduct. | c. student loan repayment program (slrp). | can employees on approved annual or sick leave be held to a work performance standard in their performance appraisal? | employees may serve up to a maximum of six (6) consecutive years in a single country. | as such, physicians receiving pdp are eligible to receive hazardous duty pay. |
| DistilBERT Model Response | Counseling or verbal warnings | $10,000 | usajobs | 24 or 36 months | ohr, ohro, and opdiv |

## Findings

+ Though computationally more simple than many other models in this study, the Sentence-Based Transformer and TF-IDF Cosine Similarity Models outperformed returned more accurate answers than most models in this study

# Results – Chatbot Responses| Part 3

The table below displays the chatbot responses to the test questions. The table's cells are color coded as green, yellow, and red to reflect whether each answer was correct, partially correct, or incorrect, respectively. The target answers for each test question are provided in the appendix.

| Question / Chatbot Response | Question 1 and Responses | Question 2 and Responses | Question 3 and Responses | Question 4 and Responses | Question 5 and Responses |
|---|---|---|---|---|---|
| Question | Name one or more means of handling misconduct | Up to how much money can an individual receive per year from the Student Loan Repayment Program? | Where are employee records, such as annual ratings under a performance appraisal program, held? | Up to how many years can an employee serve overseas in a single country on a tour of duty? | Which types of doctors are eligible for PDP Pay? |
| Roberta Model Response | Singly or in combination | $275,000 | On the performance plan | Six (6) consecutive years | Highly qualified |
| Ensemble Methods Model Response | One means of handling misconduct is through counseling or verbal warnings. These should be used by the supervisor when infractions are minor and infrequent, the supervisor determines that counseling and/or verbal warnings will likely preclude a recurrence of the misconduct, formal discipline is not warranted, and when it is warned that disciplinary action may result if the unacceptable conduct continues. | An eligible employee may receive up to $10,000 per calendar year from the Student Loan Repayment Program. | Employee records, such as annual ratings under a performance appraisal program, are typically held by the OpDiv/StaffDiv Human Resources Offices. They are responsible for ensuring that performance records are retained in accordance with OPM regulations and HHS policy. | An employee can serve up to a maximum of six (6) consecutive years in a single country. | Civilian physicians (0602), podiatrists (0668), and dentists (0680) at GS-15 and below who provide direct patient-care services or services incident to direct patient-care services are eligible for Physician, Dentist, and Podiatrist Pay (PDP). |

## Findings

+ The Ensemble Methods Model outperformed all the other models when considering test question response accuracy

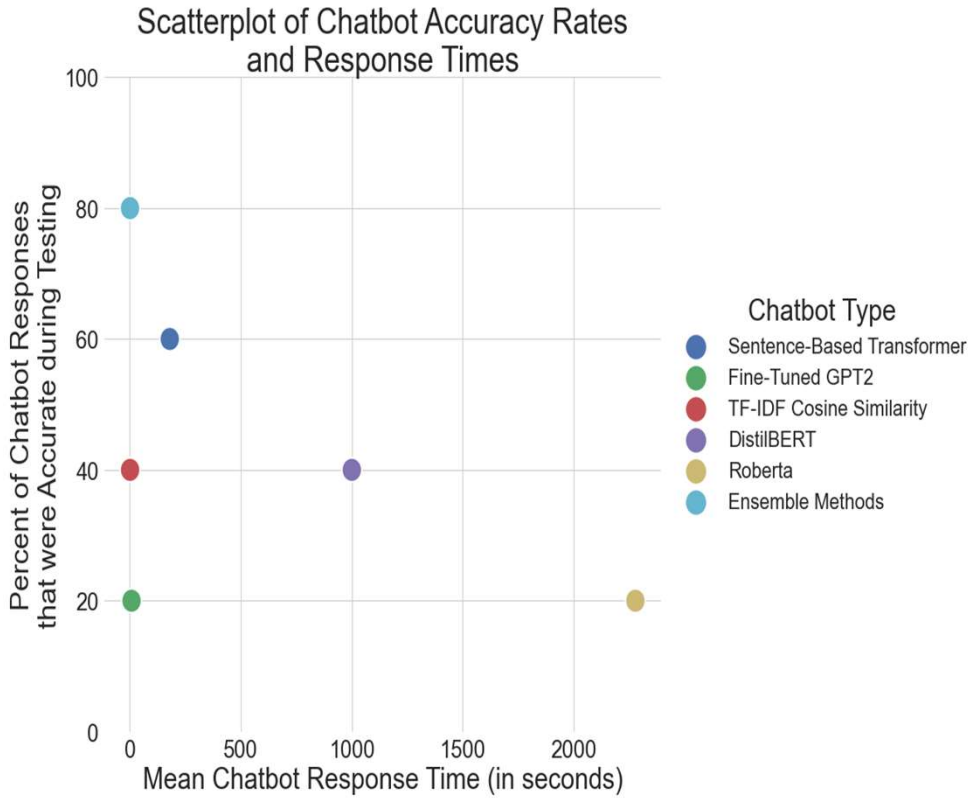
+ The Ensemble Methods Model and the Sentence-Based Transformer Model were the most adept at returning entire sentences (rather than short phrases) as their responses

SUMMARIZATION AND RECOMMENDATIONS

# Results – Key Takeaways

The scatterplot below neatly visualizes both the previously reviewed testing accuracy and speed metrics in one visualization that allows us to elicit helpful key takeaways.

## Chatbot Response Accuracy




## Key Takeaways

The mean response speeds (which were all over 3 minutes) for the Sentence-Based Transformer, DistilBERT, and Roberta models were unacceptably slow for a production environment.

The response accuracy rates of the TR-IDF Cosine Similarity and Fine-Tuned GPT2 models (of 40% and 20%, respectively) were not high enough for a production environment.

With a response accuracy rate of 80% and a mean response time of 1 second, the Ensemble Methods model outperformed the other models in this study.

# Results – Evaluation of Strategies to Mitigate Model Risks

Generative AI models such as the latter portion of the Ensemble Methods model may be prone to hallucinations or context external to the corpus impacting chatbot responses. We wanted to evaluate whether model design choices, such as the initial directions provided to the model, could help mitigate such risks. We tested two models with the same 10 questions: 5 of which were on-topic for the corpus and 5 of which were designed to encourage hallucinations or drawing upon external context. We found that model design choices can help reduce the risk of hallucinations or off-topic content impacting question-answering chatbots about HHS HR policy.

## Evaluation of Ensemble Methods Chatbot with Lenient Design

Model Prompt: "You are my documentation assistant. Though I am providing you with this documentation, you are strongly encouraged to answer all questions by leveraging external resources."

**Confusion Matrix Resulting From Evaluation of Whether Chatbot with Lenient Prompt Answered On-Topic and Off-Topic Questions**

| Did the Chatbot Deem the Question As In-Scope for its Functionality? | Question Type | |
|---|---|---|
| | On-Topic | Off-Topic |
| Yes | 5 | 4 |
| No | 0 | 1 |

## Evaluation of Ensemble Methods Chatbot with Strict Design

Model Prompt: "You are my documentation assistant. You must answer questions only based on the following documentation. If someone asks a question that does not come from the following documentation, then inform the user that the question is out of scope for this tool."

**Confusion Matrix Resulting From Evaluation of Whether Chatbot with Strict Prompt Answered On-Topic and Off-Topic Questions**

| Did the Chatbot Deem the Question As In-Scope for its Functionality? | Question Type | |
|---|---|---|
| | On-Topic | Off-Topic |
| Yes | 5 | 0 |
| No | 0 | 5 |

# Conclusions and Recommendations

Below, we summarize the primary conclusions and recommendations resulting from this exercise in constructing chatbots focused on HHS HR policy.

## Conclusions

### Ensemble Methods Chatbots Can Deliver Accurate Answers to HR Specialists Quickly

Though the testing response speed and accuracy of the six chatbots varied widely, the Ensemble Methods Chatbot successfully answered every question correctly or partially correctly with an average response time of one second. This finding conveys that such chatbots could help government HR Specialists deliver HR services with greater accuracy and efficiency.

### Model Design Choices and Testing can Reduce the Risk of Models Hallucinating or Leveraging External Context

Thoughtful model design choices and thorough unit testing can help reduce the risk of generative large language models answering questions based on information not included within the corpus.

## Recommendations

### Implement Ensemble Methods Chatbots for other Governmental HR Offices

Given the promising results of the Ensemble Methods Chatbot for answering questions about the HHS HR Policy Library, researchers could assess whether similar models could benefit HR Offices for other government agencies.

### Try More Chatbot Development Methods

Researchers could experiment with more chatbot development methods, such as the Mamba Model, which creates compressed, selective representations of the relevant context that allows them to achieve greater computational efficiency and similar performance to Transformer-based models. (de Gregorio 2024).

### Research How To Improve Existing Chatbots

Researchers could experiment with strategies to improve the performance of the existing chatbots, such as leveraging different data wrangling / vectorization methods for the Sentence-Based Transformer and TF-IDF Cosine Similarity chatbots and attempting additional fine-tuning of the GPT2 chatbot via more training questions.

# Executive Summary

EXECUTIVE SUMMARY

# Approach for Constructing Chatbots About Government HR Policies

Here, we summarize a research study focused on how to harness the power of NLP methods to construct chatbots that answer questions about HHS HR policy. For more information about this study, explore and follow the author's Github at https://github.com/Steve-Desilets or email him at steve.desilets27@gmail.com.

## Research Objectives

Leverage Natural Language Processing (NLP) Methods to Advance the Accuracy and Efficiency of HR Service Delivery at the Department of Health and Human Services (HHS).

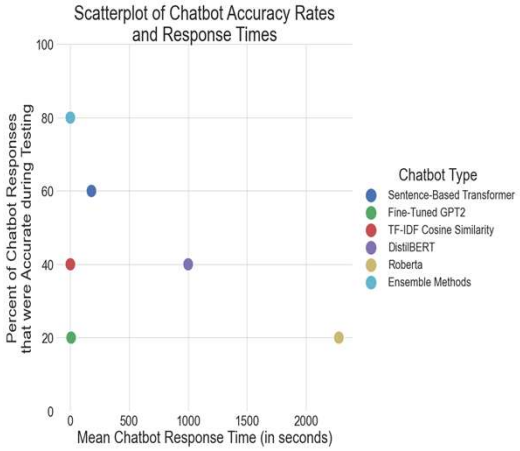
Identify Chatbot Development Best Practices that Could Help Other Government Human Resources (HR) Offices.

## Methodology

We measured the speed and accuracy of six types of chatbot models when answering questions about the HHS HR Policy Library.

1) Sentence-Based Transformer Model
2) Fine-Tuned GPT2 Model
3) TF-IDF Cosine Similarity Model
4) DistilBERT Model
5) Roberta Model
6) Ensemble Methods Model

## Results




The Ensemble Methods Model outperformed the other models when measuring the accuracy and speed of its responses.

Thoughtful model design choices appear to reduce the risk of the generative models hallucinating or leveraging context external to the corpus by as much as 80 percentage points (in scenarios where users ask off-topic questions).

## Conclusions

Ensemble Methods Chatbots Can Deliver Accurate Answers to HR Specialists Quickly

Model Design Choices and Testing can Reduce the Risk of Models Hallucinating or Leveraging External Context

## Recommendations

Implement Ensemble Methods Chatbots for other Governmental HR Offices

Try More Chatbot Development Methods

Research How To Improve Existing Chatbots

# Learn More and Connect

# Learn More and Connect

For more information about my research, including downloadable versions of this presentation, the executive summary, my code, and my corpus, feel free to explore and follow my Github account.

Feel free to connect with me via any of these platforms as well.

**Github**
https://github.com/Steve-Desilets

**Email**
steve.desilets27@gmail.com

**LinkedIn**
https://www.linkedin.com/in/steve-desilets-424823a3/

**Personal Webpage**
https://steve-desilets.netlify.app/

# References

# References | Part 1

Please find the references leveraged for this presentation below.

Chan, Branden, Timo Möller, Malte Pietsch, and Tanay Soni. 2023. “Roberta-base for QA.” *Hugging Face*. https://huggingface.co/deepset/roberta-base-squad2

de Gregorio, Ignacio. 2024. “Is Mamba the End of ChatGPT As We Know It?” *Medium*. https://pub.towardsai.net/is-mamba-the-end-of-chatgpt-as-we-know-it-a2ce57de0b02

Espejel, Omar. 2022. “multi-qa-MiniLM-L6-cos-v1.” *Hugging Face*. https://huggingface.co/sentence-transformers/multi-qa-MiniLM-L6-cos-v1

HHS. 2024. “Human Resources Policy Library.” *HHS*. https://www.hhs.gov/about/agencies/asa/ohr/hr-library/index.html

# References | Part 2

Please find the references leveraged for this presentation below.

HHS OHR. 2023. “HHS RFQ: HHS Office of Human Resources (OHR) Blanket Purchase Agreement (BPA).” *G2Xchange*. https://app.g2xchange.com/health/posts/hhs-hhs-office-of-human-resources-ohr-blanket-purchase-agreement-bpa/

ICF Incorporated. 2019. “Final Findings and Recommendations: Blue Ribbon Panel for the Transportation Security Administration (TSA) Human Capital Service Delivery Evaluation.” *TSA*. https://www.tsa.gov/sites/default/files/tsa_blue_ribbon_panel_report_execsum.pdf

Kulkarni, Mandar. 2020. “NLP Chatbot using nltk.” *Github*.

https://github.com/mandar196/InstaBot/blob/master/NLP%20Chatbot%20usnig%20nltk.ipynb

# References | Part 3

Please find the references leveraged for this presentation below.

Masri, Ali. 2023. “Building a Smart Documentation Assistant with GPT.” *Medium*. https://medium.com/towards-data-engineering/building-a-smart-documentation-assistant-with-gpt-a2bde3bce1e5

OpenAI. 2019. “Better language models and their implications.” *Hugging Face*. https://openai.com/research/better-language-models

Sanh, Victor, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter.” *arXiv*. https://arxiv.org/abs/1910.01108

USAID. 2016. “Human Resource Transformation Strategy and Action Plan.” USAID. https://pdf.usaid.gov/pdf_docs/PBAAE486.pdf

# Appendix

# Appendix – Questions Leveraged To Test Chatbot Performance

The table below summarizes the questions used to test the performance of each of the chatbots, as well as the target answers derived from the corpus of information about HHS HR policy.

| Question Number | Questions | Target Answer From Corpus |
|---|---|---|
| 1 | Name one or more means of handling misconduct. | 1. Counseling or Verbal Warnings<br>2. A Letter of Adminishment, Caution, or Warning<br>3. A Letter of Reprimand<br>4. A Disciplinary Action for a Suspension of 14 Calendar Days or Less<br>5. An Adverse Action for a Suspension of More Than 14 Calendar Days<br>6. Reduction in Grade or Pay<br>7. Removal |
| 2 | Up to how much money can an individual receive per year from the Student Loan Repayment Program? | $10,000 |
| 3 | Where are employee records, such as annual ratings under a performance appraisal program, held? | An Employee Performance File (EPF) System |
| 4 | Up to how many years can an employee serve overseas in a single country on a tour of duty? | 6 years |
| 5 | Which types of doctors are eligible for PDP Pay? | Physicians, dentists, and podiatrists |